Volume 13 Number 21 1985 Nucleic Acids Research

The nucleotide sequence of the nitrogen-regulation gene *ntrA* of *Klebsiella pneumoniae* and comparison with conserved features in bacterial RNA polymerase sigma factors

M.J.Merrick and J.R.Gibbins

AFRC Unit of Nitrogen Fixation, University of Sussex, Brighton BN1 9RQ, UK

Received 6 September 1985; Accepted 8 October 1985

ABSTRACT

The nucleotide sequence of the Klebsiella pneumoniae ntrA gene has been determined. NtrA encodes a 53,926 Dalton acidic polypeptide; a calculated molecular weight which is significantly lower than that determined by SDS polyacrylamide gel analysis. NtrA is followed by another open-reading frame (orf) of at least 75 amino acids. In the spacer region between ntrA and orf there are no apparent transcription termination or promoter sequences and therefore orf may be co-transcribed with ntrA.

Previous authors have proposed that NtrA could act as an RNA polymerase sigma factor but the NtrA amino acid sequence does not show a high level of homology to any known sigma factor. However analysis of sequences of five sigma factors from E. coli and B. subtilis has identified two conserved sequences at the C-terminal end of all these polypeptides. These sequences resemble those found in known site-specific DNA-binding domains and may be involved in recognition of conserved -35 and -10 promoter sequences. A similar pair of sequences is present at the C-terminus of NtrA and could play a role in recognition of ntr-activatable promoters.

INTRODUCTION

In Klebsiella pneumoniae expression of the nitrogen fixation (nif) genes and a number of other genes involved in nitrogen assimilation (e.g. glnA) is regulated by the nitrogen regulation (ntr) system which comprises three genes, ntrA, ntrB and ntrC (1-7). The ntrBC genes are part of a complex operon glnA-ntrBC (8,9) and the ntrA gene is linked to argG (1). A homologous nitrogen control system is present in Salmonella typhimurium (10,11), Escherichia coli (12,13), and Klebsiella aerogenes (14) but in E. coli and K. aerogenes ntrA, B and C are designated glnF, glnL and glnG respectively.

The ntrA product (NtrA) is required together with either

the ntrC product (NtrC) or the nifA product (NifA) for transcription initiation from ntr-activatable promoters, e.g. glnA or nif. Nucleotide sequence analysis of these promoters has revealed an atypical consensus sequence CTGGCACN$_5$TTGCA between positions -27 and -11 (9,15-18) rather than the characteristic sequences at -35 and -10 found in most bacterial promoters (19). Recognition of these promoters is likely to require modification of RNA polymerase and it has been proposed that NtrA (perhaps with NtrC or NifA) may behave as an alternative sigma factor to allow transcription initiation at these promoters (6,15). Ntr-regulated promoters with this characteristic consensus have also been identified in Rhizobium nif genes (20,21), Azotobacter nif genes (22,23) and the xylABC operon of Pseudomonas putida (24).

The ntrA gene of K. pneumoniae has been cloned (6,7) and its product identified as a 75 kDal acidic polypeptide (7). We have now sequenced this gene and compared the predicted amino acid sequence of NtrA with that of other known RNA polymerase sigma factors.

MATERIALS AND METHODS

Cloning and DNA sequencing

Restriction enzymes and DNA-modifying enzymes were obtained from commercial sources and used according to the manufacturers' instructions. The sequencing strategy was based on a detailed restriction map of the 1.9 kb ClaI fragment carrying ntrA which had been determined previously (7). Dideoxy sequencing reactions were carried out using defined restriction fragments cloned in M13 mp8, mp9 and mp11 vectors (25) with [α^{35}S]-dATP as the labelled nucleotide (26). Starting material for construction of M13 clones was derived from plasmids pMM17 and pBS1 (7) and pMM29 (Fig. 1). Plasmid pMM29 carries a 359 bp EcoRI-PvuII fragment from pMM17 cloned between the EcoRI and SmaI sites of the translational fusion vector pMC1403 (27). The resultant ntrA-lacZ fusion is in frame as judged by expression of β-galactosidase.

Computer analysis of sequence data employed programs developed by R. Staden (28,29).

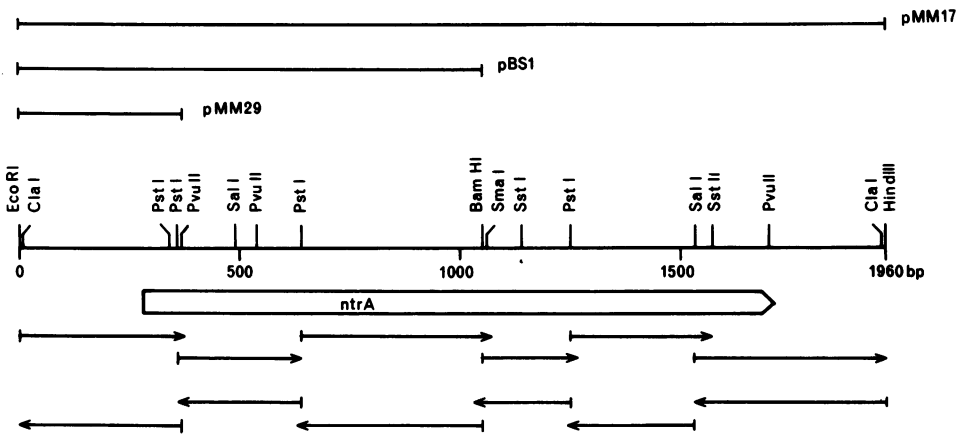

Fig. 1. Restriction map and nucleotide sequencing strategy for the ClaI fragment carrying ntrA from pMM17. The extent of fragments in two derivative ntrA-lacZ translational fusion plasmids pMM29 and pBS1, is also shown. Arrows below the map show the extent of sequence determined from each M13 clone.

RESULTS

The nucleotide sequence of ntrA

The sequence of the 1966 nucleotide EcoRI-HindIII fragment from pMM17 is shown in Fig. 2. This sequence was determined with contiguous overlapping readings for both strands and any single base position averaged more than 5 readings.

The direction of transcription of ntrA had been established from previous work (6,7) as being from the EcoRI site towards the HindIII site in pMM17. An open reading frame (orf) of 1431 bp, from positions 280 to 1710, was identified with two potential AUG codons at positions 280 and 322. Of these two codons the former was chosen as the most likely initiation codon on the basis that only it is preceded by a potential ribosome binding (SD) sequence. The predicted size of the translation product from this orf is 53,926 Daltons. This molecular weight is significantly less than that of 75-76 kDal determined by SDS polyacrylamide gel electrophoresis (PAGE) for the ntrA products of K. pneumoniae (7) and E. coli (13).

In order to confirm that the gene does terminate 250 bp before the ClaI site, plasmid pMM17 was linearised by

```
GAATTCTCATGTTTGACAGCTTATCATCGATATCAAACGCATTATTGAACACCTGCCGCGACAGCGTCTTGGCGTCCTGATAACCGACCA
      10        20        30        40        50        60        70        80        90

TAACGTCAGGGAAACGCTGGCCGTTTGCGAGCGTGCTTATATCGTCAGCCAGGGCCACCTGATAGCCCACGGTACGCCGCAGCAAATCCT
     100       110       120       130       140       150       160       170       180

TGAAGACGAGCAGGTTAAGCGCGTGTATCTTGGGGAAGACTTCAGACTCTGATAGGGTAGAGGTTACAGACGTTTTAGCCGGAGATATTG
     190       200       210       220       230       240       250       260       270

          M  K  Q  G  L  Q  L  R  L  S  Q  Q  L  A  M  T  P  Q  L  Q  Q  A  I  R  L  L  Q   27
GCCCTGAATATGAAGCAAGGTTTGCAATTAAGGCTAAGCCAACAGCTTGCCATGACGCCACAACTGCAGCAGGCGATTCGTCTACTGCAG
      280       290       300       310       320       330       340       350       360

 L  S  T  L  E  L  Q  Q  E  L  Q  Q  A  L  D  S  N  P  L  L  E  Q  T  D  L  H  D  E  V  E   57
CTGTCCACGTTAGAACTCCAGCAAGAACTCCAGCAGGCGCTGGACAGCAACCCGTTGCTGGAGCAAACCGATCTTCACGATGAGGTAGAA
      370       380       390       400       410       420       430       440       450

 T  K  E  A  E  D  R  E  S  L  D  T  V  D  A  L  E  Q  K  E  M  P  E  E  L  P  L  D  A  S   87
ACCAAAGAGGCCGAGGATCGCGAATCTCTGGATACCGTCGACGCCCTTGAGCAAAAAGAGATGCCCGAAGAGCTGCCGCTTGATGCCAGC
      460       470       480       490       500       510       520       530       540

 W  D  E  I  Y  T  A  G  T  P  S  G  N  G  V  D  Y  Q  D  D  E  L  P  V  Y  Q  G  E  T  T  117
TGGGATGAGATTTACACCGCCGGAACGCCATCAGGCAACGGCGTCGATTACCAGGATGACGAACTGCCCGTCTACCAGGGAGAGACCACG
      550       560       570       580       590       600       610       620       630

 Q  S  L  Q  D  Y  L  M  W  Q  V  E  L  T  P  F  T  D  T  D  R  A  I  A  T  S  I  V  D  A  147
CAAAGCCTGCAGGATTACCTGATGTGGCAGGTTGAACTTACGCCATTCACCGATACCGATCGCGCCATCGCGACCTCTATCGTCGATGCC
      640       650       660       670       680       690       700       710       720

 V  D  D  T  G  Y  L  T  I  S  V  E  D  I  V  E  S  I  G  D  D  E  I  G  L  E  E  V  E  A  177
GTTGATGATACCGGCTACCTGACGATTTCTGTCGAAGACATCGTGGAAAGTATTGGCGACGATGAAATCGGACTTGAAGAAGTTGAAGCG
      730       740       750       760       770       780       790       800       810

 V  L  K  R  I  Q  R  F  D  P  V  G  V  A  A  K  D  L  R  D  C  L  L  V  Q  L  S  Q  F  A  207
GTTCTCAAGCGCATTCAGCGTTTCGACCCCGTCGGCGTGGCGGCAAAAGATTTGCGTGATTGCCTGCTGGTTCAGCTTTCACAGTTTGCC
      820       830       840       850       860       870       880       890       900

 K  E  T  P  W  I  E  E  A  R  L  I  I  S  D  H  L  D  L  L  A  N  H  D  F  R  S  L  M  R  237
AAAGAGACGCCGTGGATTGAAGAAGCCCGCCTGATCATCAGCGATCATCTCGATCTGCTGGCCAACCACGACTTCCGCAGCCTGATGCGC
      910       920       930       940       950       960       970       980       990

 V  T  R  L  K  E  E  V  L  K  E  A  V  N  L  I  Q  S  L  D  P  R  P  G  Q  S  I  Q  T  G  267
GTAACCCGCCTCAAAGAAGAAGTGTTAAAGGAAGCGGTAAATCTGATCCAATCGCTGGATCCGCGCCCGGGACAGTCGATCCAAACTGGC
     1000      1010      1020      1030      1040      1050      1060      1070      1080

 E  P  E  Y  V  I  P  D  V  L  V  R  K  V  N  D  R  W  V  V  E  L  N  S  D  S  L  P  R  L  297
GAGCCAGAATATGTCATTCCTGACGTTCTGGTGCGTAAAGTCAACGATCGTTGGGTGGTTGAGCTCAATTCAGATAGCCTTCCGCGCCTG
     1090      1100      1110      1120      1130      1140      1150      1160      1170

 K  I  N  Q  Q  Y  A  A  M  G  N  S  T  R  N  D  A  D  G  Q  F  I  R  S  N  L  Q  E  A  R  327
AAGATCAATCAGCAGTATGCCGCTATGGGTAACAGCACGCGCAATGACGCTGACGGCCAGTTTATCCGTAGCAACCTGCAGGAAGCGCGC
     1180      1190      1200      1210      1220      1230      1240      1250      1260

 W  L  I  K  S  L  E  S  R  N  D  T  L  L  R  V  S  R  C  I  V  E  Q  Q  Q  A  F  F  E  Q  357
TGGCTGATCAAGAGCCTGGAGAGCCGCAACGACACCCTGCTGCGCGTCAGCCGCTGTATCGTCGAACAGCAGCAGGCTTTTTTTGAACAA
     1270      1280      1290      1300      1310      1320      1330      1340      1350

 G  E  E  F  M  K  P  M  V  L  A  D  I  A  Q  A  V  E  M  H  E  S  T  I  S  R  V  T  T  Q  387
GGTGAAGAGTTTATGAAACCGATGGTGCTGGCGGATATCGCTCAGGCCGTCGAAATGCATGAATCCACTATTTCACGCGTTACTACGCAA
     1360      1370      1380      1390      1400      1410      1420      1430      1440

 K  Y  L  H  S  P  R  G  I  F  E  L  K  Y  F  F  S  S  H  V  N  T  E  G  G  G  E  A  S  S  417
AAATACCTGCACAGTCCACGCGGTATTTTTGAGCTGAAGTATTTCTTCTCCAGCCATGTGAACACCGAAGGCGGCGGCGAAGCATCGTCG
     1450      1460      1470      1480      1490      1500      1510      1520      1530

 T  A  I  R  A  L  V  K  K  L  I  A  A  E  N  P  A  K  P  L  S  D  S  K  L  T  T  M  L  S  447
ACGGCCATTCGCGCGCTGGTGAAGAAATTAATCGCCGCGGAAAACCCCGCGAAACCATTGAGTGATAGTAAGCTCACCACCATGCTATCC
     1540      1550      1560      1570      1580      1590      1600      1610      1620

 D  Q  G  I  M  V  A  R  R  T  V  A  K  Y  R  E  S  L  S  I  P  P  S  N  Q  R  K  Q  L  V  477
GATCAGGGTATTATGGTGGCACGGCGAACCGTTGCTAAGTACCGAGAGTCTTTATCCATTCCGCCGTCAAATCAGCGCAAGCAGCTGGTT
     1630      1640      1650      1660      1670      1680      1690      1700      1710

 *                    M  Q  L  N  I  T  G  H  N  V  E  I  T  P  A  M  R  D  F  V  T  A
TGACCCAACCGATAAGGAAGACACTATGCAGCTCAACATTACAGGACACAACGTCGAGATAACCCCCGCGATGCGCGATTTCGTTACCGC
     1720      1730      1740      1750      1760      1770      1780      1790      1800

  K  F  S  K  L  E  Q  F  F  D  R  I  N  Q  V  Y  I  V  L  K  V  E  K  V  T  Q  I  A  D  A
GAAGTTCAGCAAACTGGAGCAGTTTTTCGACAGGATCAACCAGGTCTACATTGTGTTAAAAGTGGAGAAGGTGACGCAAATTGCGGACGC
     1810      1820      1830      1840      1850      1860      1870      1880      1890

  N  L  H  V  N  G  G  E  I  H  A  S  A  E  G  Q  D  M  Y  A  A  I  D  K  L  *
CAATCTGCATGTCAACGGTGGCGAAATCCATGCCAGTGCGGAAGGCCAAGATATGTATGCTGCTATCGATAAGCTTTAATGCGGTAGTTA
     1900      1910      1920      1930      1940      1950      1960      1970      1980
```

restriction at the unique HindIII site and subjected to Bal31 digestion in order to remove approximately 200 bp from the ClaI fragment. Suitably sized deletions were selected by screening plasmid digests and the endpoints of these deletions were subsequently determined after cloning in mp8. Two such deletions pMM31 and pMM32 terminated at positions 1759 and 1776 and had therefore lost 196 bp and 179 bp respectively of the ClaI fragment. These deleted plasmids still synthesised a 75 kDal polypeptide when used as templates in an S-30 *in vitro* transcription/translation system (Fig. 3).

The predicted amino acid composition of NtrA is given in Table 1. The protein has a significantly greater than average proportion of acidic and acid amide residues which is consistent with previous studies which estimated the pI of NtrA as <5.0 (7).

The Analyseq program of Staden (29) was used to identify a potential transcription initiation site for *ntrA* with a -10 sequence TATCTT at position 206 and a -35 sequence TTGAAG at position 180. These two sequences are separated by 20 bp which is slightly greater than that of 16-18 bp found in most promoters (19). The proposed *ntrA* promoter region was screened for sequences homologous to the consensus for *ntr*-activatable promoters (9, 15) and for a sequence homologous to the consensus binding site for NtrC in *ntr*-repressible promoters (9), using weight matrices developed by R. Dixon and implemented with the Analyseq program of Staden. No good fit was found for either consensus sequence.

Analysis of the sequence downstream of *ntrA* showed no classical *rho*-independent terminator structure (GC-rich stem loop followed by several T's) and a second *orf* was identified starting at position 1736 and terminating within the vector sequences (Fig. 2). This *orf* would encode a prematurely terminated polypeptide of mol. wt. 8648 Dal. The second *orf* has a good SD

Fig. 2. Nucleotide sequence of the 1966 bp EcoRI-HindIII fragment from pMM17 (see Fig. 1). Nucleotides are numbered from the first base of the EcoRI site. Proposed ribosome binding sites (positions 270, 1725) and *ntrA* promoter sequences (positions 180 and 206) are underlined. The sequence is extended 14 bp beyond the HindIII site to indicate premature termination of ORF2 in the vector sequences. Numbers at the right hand side of the figure indicate numbers of the amino acid residues in NtrA.

Table 1. Amino acid composition (mole %) of NtrA compared with that of other proteins.

Amino acids	K. pneumoniae ntrA	E. coli rpoD (ref. 30)	E. coli nusA (ref. 32)	B. subtilis rpoD (ref. 31)	Average protein (ref. 48)
Acidic (D+E)	16.7	20.4	18.6	21.1	11.5
Acid + acid amide (D+N+E+Q)	27.6	28.4	26.1	28.0	19.8
Basic (K+R+H)	11.2	14.5	12.3	16.0	13.5
Hydrophobic (L+V+I+M)	26.7	25.4	27.0	24.9	20.2
Aromatic (F+Y+W)	5.4	5.2	5.0	6.0	8.3
Charged (D+E+K+R+H)	27.9	34.9	30.9	37.1	25.1
Aliphatic (A+G)	10.9	12.1	17.6	11.3	16.9
Hydroxyl (S+T)	13.2	10.8	8.9	9.7	13.1
Molecular weight (kDal)					
(i) from DNA sequence	54	70	55	43	
(ii) SDS-PAGE	75	82-90	69	55	
pI	<5.0	4.8-5.1	4.6	-	

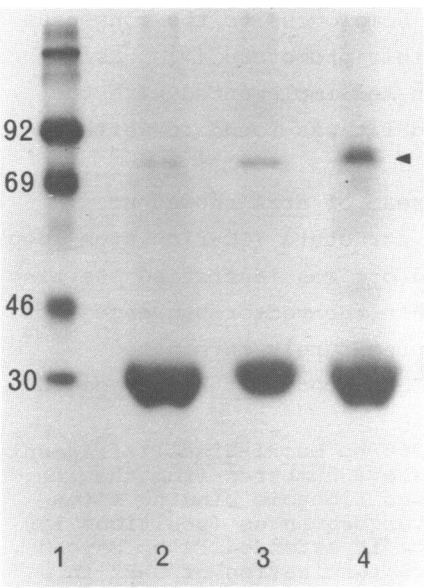

Fig. 3. SDS-polyacrylamide electrophoresis of in vitro transcription translation products synthesised in an E. coli S-30 system. Track 1 - [^{14}C] mol.wt. markers, Track 2 - pMM17, Track 3- pMM32, Track 4 - pMM31. The arrow indicates the ntrA product of 75 kDal. The 30 kDal polypeptide is the bla product from the pBR327 vector.

sequence at position 1725 and is potentially a second gene in the ntrA operon.

DISCUSSION

The nucleotide sequence of ntrA

The size of the ntrA product predicted from the DNA sequence (53.9 kDal) is significantly less than that observed by SDS-PAGE (75-76 kDal). Similar aberrant mobilities on gels have been observed for the products of E. coli rpoD (30), B. subtilis rpoD (31) and E. coli nusA (32) (see Table 1). All of these proteins are highly acidic and this anomalous behaviour has been attributed to their unusually high negative charge (30). The ntrA coding region was mapped previously by Tn5 mutagenesis (6) and this approach gave a minimal estimate of 1700±200 bp which is in reasonable agreement with our sequence.

Previous studies of ntrA expression both in K. pneumoniae and E. coli indicated that the gene was constitutively expressed at a low level and that expression was not affected by mutations in ntrA, B or C (6,7,33). Our identification of a potential -35, -10 consensus promoter sequence and the absence of sequences characteristic of ntr-regulated promoters is consistent with these earlier observations. Likewise the presence of a poor ribosome binding site is consistent with the low level of expression observed from ntrA-lacZ translational fusions.

The expression of the vector-encoded tetracycline resistance (tet) gene in two independently constructed ntrA$^+$ plasmids pMM17 (7) and pFB71 (6) was originally attributed to a fortuitous transcriptional fusion of tet to the ntrA promoter. Our sequence supports this proposition and indicates that transcription initiated at the ntrA promoter could continue through ntrA, through the prematurely terminated second orf and into the tet gene.

The identification of a second orf apparently in the same operon as ntrA raises the question of the function of this second gene. If these genes are co-transcribed it is possible that the functions of the two gene products are related. Relatively little genetic analysis of the ntrA region has been undertaken (10,13) and further experiments to determine the

```
Ec RpoD    (111)    LTREGEIDIAKRIED-G * AKKEMVEANLRLVISIAKKYTNR..
Bs RpoD    (114)    LSAKEEIAYAQKIEE-GDEESKRRLAEANLRLVVSIAKRYVGR..
Bs SpoIIG   (39)    LSKDEEQVLLMKLPN-GDQAARAILIERNLRLVVYIARKFENT..
Ec HtpR     (29)    LSADEERALAEKLHYHGDLEAAKTLILSHLRFVVHIARNYAGY..

Ec RpoD             GLQFLDLIQEGNIGLMKAVDKFEYRKGYKFSTYATWWIRQAI
Bs RpoD             GMLFLDLIHEGNMGLMKAVEKFDYRRGYKFSTYATWWIRQAI
Bs SpoIIG           GINIEDLISIGTIGLIKAVNTFNPEKKIKLATYASRCIENEI
Ec HtpR             GLPQADLIQEGNIGLMKAVRRFNPEVGVRLVSFAVHWIKAEI
```

Fig. 4. Alignment of homologous regions in E. coli RpoD and HtpR, and B. subtilis RpoD and SpoIIG. Amino acid residues are defined by the standard one letter code. Figures in brackets refer to the position of the first amino acid shown in each protein. The * in E. coli RpoD indicates the position of a 248 residue insert not present in the other sequences. Residues conserved in two or more proteins are underlined and I, L and V are considered as conservative substitutions.

phenotype of mutations in this region will be of interest.

Comparison of the NtrA sequence with other sigma proteins

A number of genes which encode known RNA polymerase sigma factors have been cloned and sequenced. These include E. coli rpoD (30) and htpR (34,35), B. subtilis rpoD (31) and spoIIG (36) and B. subtilis bacteriophage SPO1 genes 28 (gp28) (37) and 34 (gp34) (38). The sequence is also available for E. coli NusA which, although not a sigma factor, is known to interact with RNA polymerase thereby modifying both transcription termination and antitermination (32). Previous authors have compared various of these sequences and some regions of homology have been identified. E. coli RpoD and B. subtilis RpoD are highly homologous (31) and a region of 84 residues is conserved between these two proteins and E. coli HtpR and B. subtilis SpoIIG (31,34,36) (see Fig. 4). This homology is interrupted in E. coli RpoD by an insert of 248 residues not present in the other sequences (31) and consequently the conservation of the first twenty residues shown in Fig. 4 was not recognised in previous comparisons (34,36). No amino acid sequences homologous to this conserved region are present in Gp28, Gp34, or NusA and no comparable homology could be identified in NtrA.

By further sequence analysis we have now identified two other regions of homology present in E. coli RpoD, HtpR, NusA; B. subtilis RpoD, SpoIIG and SPO1 Gp34 (Fig. 5A). These regions

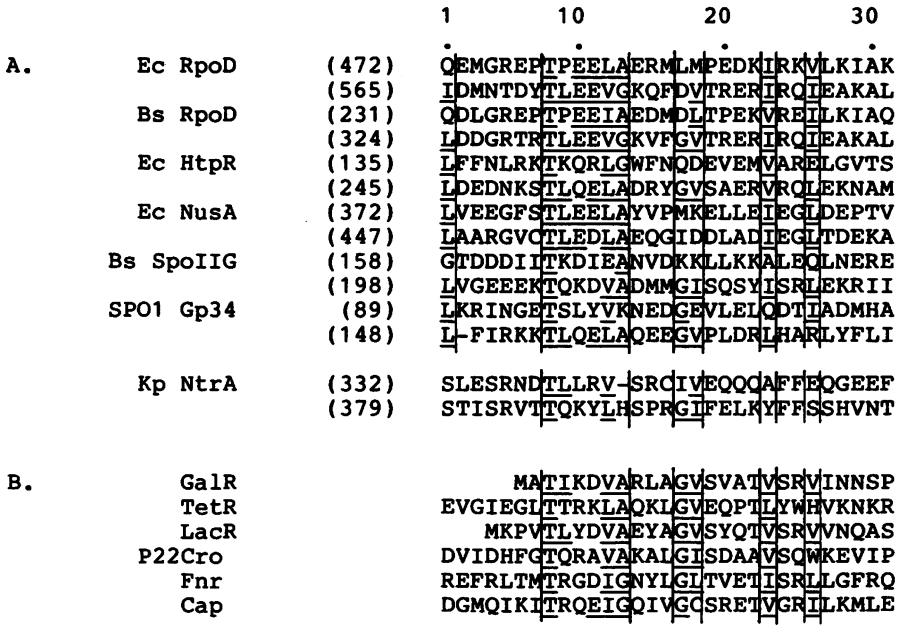

Fig. 5. A. Homologous regions in E. coli RpoD, HtpR, NusA; B. subtilis RpoD, SpoIIG, SPO1 Gp34 and K. pneumoniae NtrA. Amino acid residues are defined by the standard one letter code and conserved residues are underlined (I, L and V are considered as conservative substitutions). Figures in brackets refer to the position in the protein of the first amino acid in each sequence. Figures above the sequences are for reference purposes (see text). B. Comparative sequences from known site-specific DNA binding proteins aligned according to ref. 39.

are of particular interest as they include features which have been proposed to be characteristic of site-specific DNA binding proteins (39). In all cases these homologous regions are at the C-terminal end of the protein and are separated by a region of 40 to 100 residues. Two of these regions, residues 331 to 352 in B. subtilis RpoD and residues 253 to 272 in E. coli HtpR, have previously been identified as possible DNA-binding domains (31,34).

The homologous regions have a number of features in common which identify them as potential DNA binding sites (Fig. 5A). (i) Conservation of hydrophobic residues at positions 1, 12, 18, 23 and 26; (ii) an invariant threonine at position 8; (iii) conserved polar residues at positions 10 and 11; (iv) a

		10		10
		•		•
Ec RpoD	(479)	TPEELAERMLM	(572)	TLEEVGKQFDV
Bs RpoD	(238)	TPEEIAEDMDL	(331)	TLEEVGKVFGV
Ec HtpR	(142)	TKQRLGWFNQD	(252)	TLQELADRVGV
Bs SpoIIG	(165)	TKDIEANVDKK	(205)	TQKDVADMMGI
SPO1 Gp34	(96)	TSLYVKNEDGE	(155)	TLQELAQEEGV
Ec NusA	(379)	TLEELAYVPMK	(454)	TLEDLAEQGID
Kp NtrA	(339)	TLLRV-SRCIV	(386)	TQKYLHSPRGI

Fig. 6. Selected sequences from Fig. 5 aligned to show the higher degree of homology present in the more C-terminal of the two homologous regions in each protein. Other details are as in Fig. 5.

conserved alanine or glycine at position 13. For comparative purposes the relevant sequences from a number of known site-specific DNA binding proteins are shown in Fig. 5B.

When intergenic comparisons are made for the two regions in each protein the most C-terminal of the two regions shows the greatest degree of conservation (Fig.6), particularly with respect to the conserved glycine and adjacent hydrophobic residue at positions 17 and 18. These two residues are characteristic of the tight turn between the two α-helices involved in DNA binding but neither is apparently invariant (39).

When the NtrA sequence was searched for comparable sequences to those described above, two potential regions (residues 332 to 361 and 379 to 409) were identified (Fig. 5A). In each case the degree of homology is less than that seen in the other six sequences. Analysis of the SPO1 Gp28 sequence failed to identify any regions homologous to the sequences in Fig. 5A.

Role of potential DNA-binding domains?

The identification of two quite closely linked (40-100 residues apart) potential DNA-binding sites in five functionally related proteins raises the question of the possible functions of such sites. Whilst RNA polymerase core enzyme alone possesses catalytic activity, it is the sigma subunit which is required for promoter recognition. Isolated sigma factor from E. coli and B. subtilis has been shown to bind to supercoiled DNA, but not in a site-specific fashion (40,41). However, cross-linking studies with RNA polymerase holoenzyme have shown that sigma can cross-

link to the promoter (42) suggesting that promoter selection is probably dictated by direct interaction of sigma factor, as part of the holoenzyme complex, with specific nucleotides in the promoter. In E. coli two different sigma factors are known (30,43) and in B. subtilis five different sigma factors have been identified as well as two phage SPO1-encoded sigmas, and each apparently recognises specific -35 and -10 promoter consensus sequences (44).

Two models for the role of sigma factors in promoter selection have been proposed (45). In the 'core-conformation' model each sigma contacts a characteristic -10 sequence and induces a core-conformation that favours a particular canonical sequence at the -35 position. In the direct-contact model, favoured by Losick and Pero (45), sigma contacts both the -35 and -10 regions, either simultaneously or sequentially, during the formation of the RNA polymerase-promoter complex. Such a model requires each sigma factor to have two domains which will mediate -35 and -10 recognition. These domains would not necessarily be expected to resemble closely the structures found in site-specific DNA binding proteins as the nature both of the DNA sequence recognised and the type of protein DNA interaction are probably different in sigma factors. Nevertheless, the identification of two domains which resemble the consensus DNA binding site of site-specific DNA binding proteins suggests that these regions may be those which are concerned with recognition of the -35 and -10 promoter sequences. An alternative, but perhaps less likely, role for these conserved sequences is that of a protein/protein recognition domain, i.e. a region of interaction with one or more subunits of RNA polymerase core enzyme.

As described earlier, ntr-activatable promoters have an entirely different consensus, both in sequence and spacing, from that found in other prokaryotic promoters. Hence, if NtrA is indeed a sigma factor it might be expected to differ from other sigma factors in respect of any potential DNA-recognition domains. The two regions we have identified in NtrA are similar but not distinctly homologous to the very well conserved regions in other sigma factors and like the other regions they are at the C-

terminus of the protein. It remains to be demonstrated whether these NtrA sequences or those identified in the other sigma factors do indeed play a direct role in promoter selection.

The role of the consensus sequences in NusA may differ from that in other sigma factors. However, a non-symmetrical DNA sequence (boxA) has been identified as a potential recognition site for NusA suggesting that transcription termination by an RNA polymerase/NusA complex requires interaction between NusA and the boxA sequence (46). Recent studies of transcriptional auto-regulation by NusA suggested that a NusA:chloramphenicol trans-acetylase (CAT) hybrid protein, in which CAT is fused after residue 456 of NusA, is defective in normal regulation (47) indicating that integrity of the second consensus sequence (residues 447-477) may be necessary for normal NusA function. It is notable that the two NusA DNA-recognition sequences show much greater inter-sequence homology than that seen between other intergenic pairs and in NusA the region of homology extends over 48 residues of which 19 are identical and 6 show conservative replacements.

The absence of any obvious homology between SPO1 Gp28 and other sigma factors is surprising. However until the nature of the interaction between sigma factors and RNA polymerase core enzyme, and between RNA polymerase and promoter sequences, is better understood it may not be possible to predict potentially homologous domains from the primary amino acid sequence of sigma factors.

In this paper we believe we have identified some potentially interesting sequences in a number of sigma factors and in NtrA. It remains for future studies to determine whether these sequences are functionally homologous and structurally significant in the mode of action of RNA polymerase and associated sigma factors.

ACKNOWLEDGEMENTS

We would like to thank Ray Dixon, Martin Drummond and Martin Buck for much advice and valuable discussion. We also thank Brenda Hall for typing the manuscript.

Since this paper was submitted for publication a comparable analysis of bacterial sigma factors has been published by Stragier et al., (FEBS Letts. 187, 11-15, 1985). These authors analyse the E. coli RpoD and HtpR and B. subtilis RpoD and SpoIIG sequences and identify the same potential DNA binding sites as those described in this paper.

REFERENCES

1. Leonardo, J.M. and Goldberg, R.B. (1980) J.Bacteriol., 142, 99-110.
2. Espin, G., Alvarez-Morales, A. and Merrick, M. (1981) Mol. Gen.Genet., 184, 213-217.
3. Espin, G., Alvarez-Morales, A., Cannon, F., Dixon, R. and Merrick, M. (1982) Mol.Gen.Genet., 186, 518-524.
4. Merrick, M. (1983) EMBO J., 2, 39-44.
5. De Bruijn, F.J. and Ausubel, F.M. (1981) Mol.Gen.Genet., 183, 289-297.
6. De Bruijn, F.J. and Ausubel, F.M. (1983) Mol.Gen.Genet., 192, 342-353.
7. Merrick, M.J. and Stewart, W.D.P. (1985) Gene, 35, 297-303.
8. Alvarez-Morales, A., Dixon, R. and Merrick, M. (1984) EMBO J. 3, 501-507.
9. Dixon, R. (1984) Nucleic Acids Res., 12, 7811-7830.
10. Garcia, E., Bancroft, S., Rhee, S.G. and Kustu, S. (1977) Proc.Natl.Acad.Sci.USA, 74, 1662-1666.
11. McFarland, N., McCarter, L., Artz, S. and Kustu, S. (1981) Proc.Natl.Acad.Sci.USA, 78, 2135-2139.
12. Pahel, G., Rothstein, D.M. and Magasanik, B. (1982) J.Bacteriol., 150, 202-213.
13. Magasanik, B. (1982) Ann.Rev.Genet., 16, 135-168.
14. Rothman, N., Rothstein, D., Foor, F. and Magasanik, B. (1982) J.Bacteriol., 150, 665-671.
15. Beynon, J., Cannon, M., Buchanan-Wollaston, V. and Cannon, F. (1983) Nature, 301, 302-307.
16. Drummond, M., Clements, J., Merrick, M.J. and Dixon, R. (1983) Nature, 301, 302-307.
17. Ow, D.W., Sundaresan, V., Rothstein, D.M., Brown, S.E. and Ausubel, F.M. (1983) Proc.Natl.Acad.Sci.USA, 80, 2524-2528.
18. Dixon, R.A. (1984) J.Gen.Microbiol., 130, 2745-2755.
19. Hawley, D.K. and McClure, W.R. (1983) Nucleic Acids Res., 11, 2237-2255.
20. Ausubel, F.M. (1984) Cell, 37, 5-6.
21. Alvarez-Morales, A. and Hennecke, H. (1985) Mol.Gen.Genet., 199, 306-314.
22. Brigle, K.E., Newton, W.E. and Dean, D.R. (1985) Gene, in press.
23. Robson, R., Jones, R., Woodley, P. and Evans, D. (1985) in Proceedings of 6th International Symposium on Nitrogen Fixation, Evans, H.J. and Newton, W.E. Eds., Nijhoff/Junk, Boston (in press).
24. Dixon, R. (personal communication).
25. Messing, J. (1983) in Methods in Enzymology, Wu, R., Grossman, L. and Moldave, K. Eds., Vol. 101, pp.20-77. Academic Press, New York.
26. Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) Proc.Natl. Acad.Sci.USA, 80, 3963-3965.

27. Casadaban, M.J., Chou, J. and Cohen, S.N. (1980) J.Bacteriol. 143, 971-980.
28. Staden, R. (1982) Nucleic Acids Res., 10, 4731-4751.
29. Staden, R. (1984) Nucleic Acids Res., 12, 521-538.
30. Burton, Z., Burgess, R.R., Lin, J., Moore, D., Holder, S. and Gross, C.A. (1981) Nucleic Acids Res., 9, 2889-2903.
31. Gitt, M.A., Wang, L-F. and Doi, R.H. (1985) J.Biol.Chem., 260, 7178-7185.
32. Ishii, S., Ihara, M., Maekawa, T., Nakamura, Y., Uchida, H. and Imamoto, F. (1984) Nucleic Acids Res., 12, 3333-3342.
33. Castano, I. and Bastarrachea, F. (1984) Mol.Gen.Genet., 195, 228-233.
34. Landick, R., Vaughn, V., Lau, E.T., VanBogelen, R.A., Erickson, J.W. and Neidhardt, F.C. (1984) Cell, 38, 175-182.
35. Yura, T., Tobe, T., Ito, K. and Osawa, T. (1984) Proc.Natl. Acad.Sci.USA, 81, 6803-6807.
36. Stragier, P., Bouvier, J., Bonamy, C. and Szulmajster, J. (1984) Nature, 312, 376-378.
37. Costanzo, M. and Pero, J. (1983) Proc.Natl.Acad.Sci.USA, 80, 1236-1240.
38. Costanzo, M., Brzustowicz, L., Hannett, N. and Pero, J. (1984) J.Mol.Biol., 180, 533-547.
39. Pabo, C.O. and Sauer, R.T. (1984) Ann.Rev.Biochem., 53, 293-321.
40. Kudo, T. and Doi, R.H. (1981) J.Biol.Chem., 256, 9778-9781.
41. Kudo, T., Jaffe, D. and Doi, R.H. (1981) Mol.Gen.Genet., 181, 63-68.
42. Simpson, R.B. (1979) Cell, 18, 277-285.
43. Grossman, A.D., Erickson, J.W. and Gross, C.A. (1984) Cell, 38, 383-390.
44. Johnson, W.C., Moran, C.P. and Losick, R. (1983) Nature, 302, 800-804.
45. Losick, R. and Pero, J. (1981) Cell, 25, 582-584.
46. Friedman, D.I. and Olson, E.R. (1983) Cell, 34, 143-149.
47. Plumbridge, J.A., Dondon, J., Nakamura, Y. and Grunberg-Manago, M. (1985) Nucleic Acids Res., 13, 3371-3388.
48. Dayhoff, M.O., Hunt, L.T. and Hurst-Calderone, S. (1978) in Atlas of Protein Sequence and Structure, Dayhoff, M.O. Ed., Vol. 5, pp.363-369.